A Low Power High Throughput Architecture for Deep Network

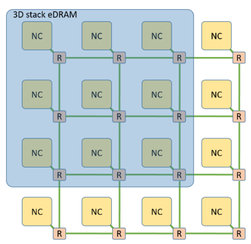
Yangjie Qi and Rasitha Fernando

Advisors: Dr. Tarek M. Taha

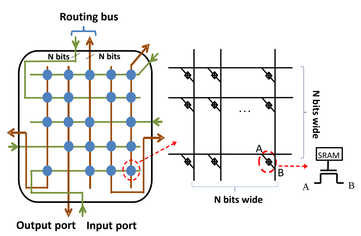
Introduction

- Deep learning is a very important set of algorithms that are now used for processing data in multiple ways. One of the key aspects of deep learning, is training. This is where the algorithm learns how to classify new data. Training is very expensive in time and energy, and thus is only done on large powerful computers.
- We have developed a novel specialized chip that can do this learning at very low power consumption. The reduced power consumption will allow this chip to be used in everyday devices like cell phones, medical devices, and robots (the range of devices is enormous), to make them much more smart as they would be able to adapt on their own.

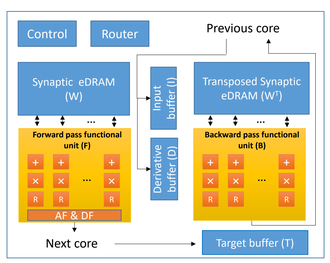
Proposed System Overview



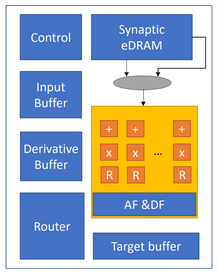
Proposed multicore system



SRAM based static routing switch

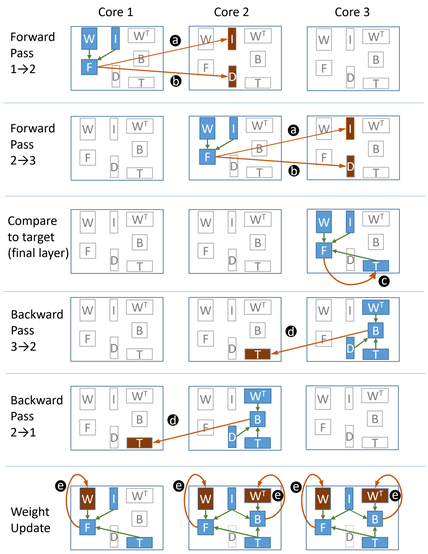


Single-Channel Dual-Memory Neural Core Design



Dual-Channel Single Memory Neural core Design

The Algorithm and Core Design



The Data Flow for Training

Dataset	Configuration
MNIST	784→400→100→10
COIL-20	1024→512→256→128→20
COIL-100	3072→1536→768→384→192→100

Network Configuration

Calculation	Equation	Circuit
a	(1) + (2)	C1
b	(3)	C1
c	(2) + (4)	C1
d	(5)	C2
e	(6)	C3 and C4

The Relationship

$$DP^{(i+1)} = Z^{(i)}W^{(i)\to(i+1)} \quad (1)$$

$$Z^{(i+1)} = f(DP^{(i+1)}) \quad (2)$$

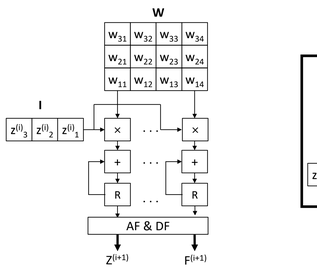
$$F^{(i+1)} = f'(DP^{(i+1)}) \quad (3)$$

$$D^{(out)} = F^{(i)} \odot (Z^{(out)} - Y) \quad (4)$$

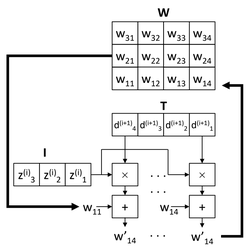
$$D^{(i)} = F^{(i)} \odot D^{(i+1)}(W^{(i)\to(i+1)})^T \quad (5)$$

$$\Delta W^{(i)\to(i+1)} = -\eta((D^{(i+1)})^T(Z^{(i)})^T) \quad (6)$$

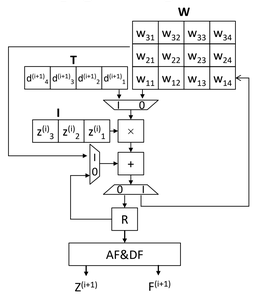
Formula for Back Propagation Algorithm



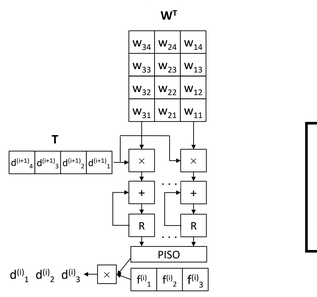
Circuits for Inference Pass (C1)



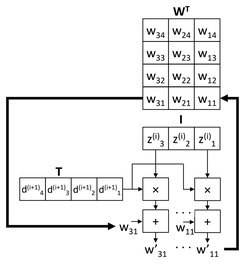
Circuits for Weight Update (C3)



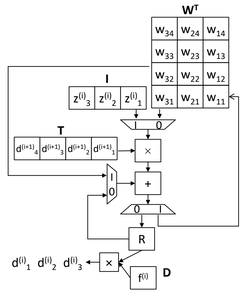
Circuits for Inference Pass and Weight Update (C5)



Circuits for Backpropagation (C2)

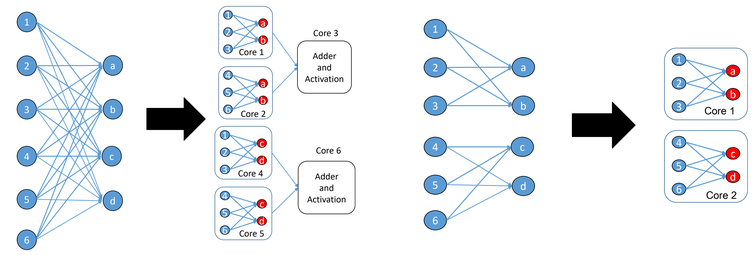


Circuits for Transposed Weight Update (C4)



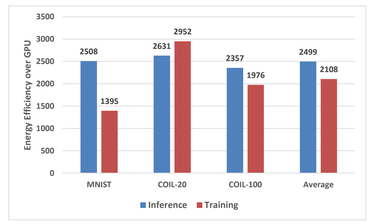
Circuits for Backpropagation and Transposed Weight Update (C6)

Network Mapping Strategy

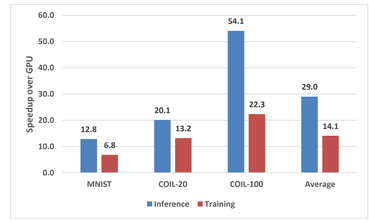


General Mapping Strategy **Proposed Mapping Strategy**

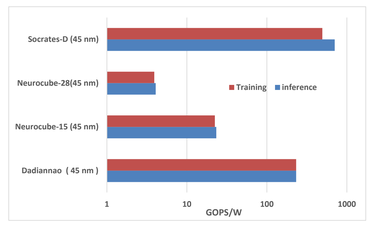
Speedup and Energy Efficiency



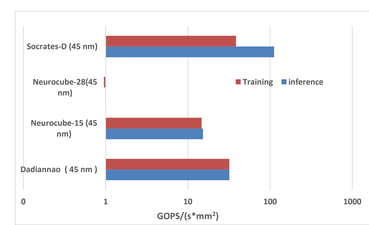
Energy Efficiency over GPU



Speedup over GPU



Power Efficiency Compare with Other Design



Computation Efficiency Compare with Other Design

Conclusion

- The system have both training and recognition (evaluation of new input) capabilities.
- The chip was about 2000x more energy efficient and about 14 times faster than an NVIDIA graphics processor for learning multiple types of data.

High Performance Computing Laboratory

Kettering Laboratories
300 College Park
Dayton, OH 45469-0232